PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G05B 5/34

A1

(11) International Publication Number: WO 99/60450

(43) International Publication Date: 25 November 1999 (25.11.99)

(21) International Application Number: PCT/US99/11259

(22) International Filing Date: 20 May 1999 (20.05.99)

(30) Priority Data:
09/084,110 21 May 1998 (21.05.98) US

(71) Applicant *(for all designated States except US)*: SMITHKLINE BEECHAM CORPORATION [US/US]; One Franklin Plaza, Philadelphia, PA 19103 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: GRELLER, Larry, D. [US/US]; 486 Reginald Lane, Collegeville, PA 19426 (US). TOBIN, Frank, L. [US/US]; 324 Crum Creek Lane, Newtown, PA 19073 (US).

(74) Agents: BAUMEISTER, Kirk et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(81) Designated States: CA, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHODS AND SYSTEMS OF IDENTIFYING EXCEPTIONAL DATA PATTERNS

(57) Abstract

A computational method for the identification of exceptional values in arrays of many sorts of intensity data is provided. The method is indifferent as to whether the intensities are experimental or computationally derived. Identification of patterns of selective expression of mRNA or protein gene products can be provided by the method of the invention.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

Methods and Systems of Identifying Exceptional Data Patterns

Field of the Invention

This invention relates to computer-based methods and systems for identification of exceptional patterns in data, such as selectively expressed genes and gene products.

Background of the Invention

The general problem of identifying exceptional patterns in data from many different sources can be viewed as an outlier identification problem. The outlier concept and statistical methods for outlier detection have an extensive literature [17-20]. Yet, what kinds of interpretations and quantitative treatments of data define an outlier remains fluid statistically and scientifically [17-20] and subjective [17].

Outlier detection problems arise in many different contexts. In the drug discovery field, intensity patterns may come from any array of intensity data derived from, for example, EST sequencing, microarray DNA hybridization, macromolecular gridding, compound assay data, molecular screening data, patient diagnostic and toxicological data. The conjunction of large-scale biology technologies, such as genomic sequencing or proteomics, and the need for new drug discovery targets has resulted in a need for more robust methods for detecting unusual expression patterns across many data sources. Thus, a need exists for useful quantitative objectivity to be brought to bear on the fundamental subjectivity of outlier detection.

Summary of the Invention

Accordingly, one aspect of the present invention is a method of identifying selectively expressed (exceptional) values in intensity data comprising analyzing statistical discordancy and gap criterion in a decision function wherein the decision function provides an overall confidence of above- or below-baseline exceptional intensity identification.

Another aspect of the invention is a method of identifying selectively expressed values in intensity data comprising:

(a) selecting intensity values from intensity data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensities exceeds a predetermined minimum;

(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of exceptional intensity;

(f) identifying the degree of overall confidence of exceptional intensity; and

(g) displaying the results of step (f) on an output device.

Another aspect of the invention is a method of detecting selective expression of gene or gene products comprising:

(a) selecting intensity values from gene product data sources, wherein the source quality weight exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensity values exceeds a predetermined minimum;

(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) determining a gap by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) applying a decision function to the statistical significance and the gap to determine an overall confidence of selective expression;

(f) identifying the degree of overall confidence of selective expression; and

(g) displaying the results of step (f) on an output device.

Yet another aspect of the invention is computer systems and computer readable media for performing the methods of the invention.

Brief Description of the Drawings

Fig. 1 diagrams simple stereotypical examples of selective expression types "up," "down," and "mixed". Intensities vs. sources from a source set are plotted in arbitrary order. Selectively expressed intensities are indicated by encircled symbols.

Fig. 2 shows separation of a largest value from the *n-1* others where x_i represents the intensities being compared in ascending order, $x_{i-1} \leq x_i$, $i = 1, \ldots, n$. The basic measures for the Dixon test, namely the distance between the largest and the next-to-largest values (gap = $x_n - x_{n-1}$) and the distance between the largest and smallest values ($x_n - x_1$), are used to calculate the separation ratio τ = gap / ($x_n - x_1$).

Fig. 3 shows discordancy statistical significance adjusted for baseline position. Synthetic intensity data vs. source for a variety of different baseline levels of intensity, {0.25, 0.5, 0.75, and 0.9} are plotted.

Fig. 4 shows how erosion of statistical confidence increases as the baseline position increases towards the allowed maximum. Erosion of statistical confidence, i.e., loss of discordancy significance from the traditional Dixon value, is plotted vs. baseline encroaching toward the allowed maximum.

Fig. 5 shows a plot of a decision function, *d*, contours for selective expression (s.e.) overall confidence.

Fig. 6, panels A and B, shows examples of synthetic intensity (abundances) vs. source (library) data for assemblies. Panel C shows source qualities.

Fig. 7 shows stereotypical examples of selective expression in real data detected by the algorithm of the invention.

Detailed Description of the Invention

The method of the invention presents robust computational algorithms that identify exceptional values in intensity data. The algorithms are well-suited for the identification of exceptional values in many sorts of intensity data, even noisy data.

The method is generally applicable to any kind of intensity data where a distinguishable data source such as tissue, cDNA library, human, non-human (such as animal, plant, viral, bacterial or other microbial) source can be associated with each intensity value (e.g., gene or protein abundance, clone, biological or chemical activity, binding strength or genetic polymorphism assessment). For example, intensity values can be obtained from genomic sequencing, EST sequencing, microarray DNA hybridization, macromolecular gridding, compound assays, molecular screening assays, patient diagnostic or toxicological data sources. Assessments of trust, reliabilities, or relevances in the sources can be used as a basis for confidence. The intensities can be experimentally determined values, computationally derived values (e.g., abundances from cDNA data), or combinations. The method is indifferent to the experimental or computational lineages of the data to be analyzed. All that is required are triples of associated elements: entity (e.g., gene, protein, clone, assay, compound, etc.), intensity, and source. Table 1 lists some exemplary contexts where the method of the invention can be applied.

TABLE 1 - Different Contexts for Application of the Selective Expression Algorithm

	Entity	Source Set comprising ...	Intensity	Typical Question Associated with the Context
1	contig, orf, assembly, gene, clone or protein	any patient, tissues or libraries of interest generally	abundance	What genes are selectively expressed?
2	assembly, gene, or protein	different patients, tissues, libraries receiving different treatments	abundance	What selectively expressed genes are associated with which tissues and specific treatments?
3	assembly, gene, or protein	same tissue type exposed to different doses of a compounds	abundance	What are the selective expression dose responses?
4	assembly, gene, or protein	same tissue type receiving a series of related treatments	abundance	What is the selective expression in response to a specific series of treatments?

5	assembly, gene, or protein	same tissue type at different times after a single treatment	abundance	What is the kinetics (i.e., time course) of selective expression?
6	assay	compounds	biological or chemical activity	Is there an assay whose intensity is selectively expressed among the compounds tested?
7	compound	genes of toxicological interest in a single tissue, e.g., liver	abundance	Is there a gene of toxicological interest selectively expressed in response to the compound?
8	compound	screening assays, chemical or biological	assay activity	Is there a selectively expressed assay activity in a particular assay in response to compound?
9	patient, assembly, gene, protein, or compound	any interpretable combination of the above	abundances or assay activities scored for comparison	What entities are selectively expressed in any interpretable source set?

As used herein, "source" means any entity which may provide an intensity, e.g., tissue or EST library for genes or gene products, biological or chemical assay for compounds. "Genes" includes genomic DNA copy number, RNA, RNA transcripts. "Gene products" include proteins and RNA transcripts. If a source is experimentally manipulated or edited in any way, e.g., a normalized or subtracted cDNA library [9-11], it should not be included in the analysis lest its pattern of expressed genes be artificially skewed. This exclusion principle can be relaxed if all the sources being compared have been manipulated in the same way.

As used herein, "source set" means any collection comprising selected sources which may be analyzed for intensity patterns.

As used herein, the term "source confidence" represents the quality, the trust, the reliability, the knowledge of error, or the relative importance that can be attributed to the intensities obtained from the source. For example, a cDNA library

sequenced in depth is a more reliable source than the same library sequenced to less depth.

As used herein, "source quality weights" represents quantitation of source confidences. Any consistent source quality weighting scheme can be used, but care must be exercised. If the weights are not faithful to the scientific reliabilities of the sources, any results dependent upon them can be improperly distorted. An edited or normalized cDNA library, for example, should be considered a low confidence source, i.e., given small weight, in a selective expression determination unless all the sources in the source set have been manipulated equivalently.

As used herein, "intensity" means a measured or calculated non-negative numerical value which is assigned to an observation, whether the observation is experimentally and/or computationally derived from data. For example, intensity could be a drug's binding affinity, a compound's activity in a screen, or a gene's abundance such as the gene product's copy number (molecules or concentration of mRNA) or amount of protein expressed. Intensity can be either an experimentally measured quantity, or less directly, a quantity which is calculated, for example, from analyses of cDNA assemblies [9, 12, 13]. For each source, the intensities may be scaled by a suitable norm, e.g., the maximum intensity, observed in that source. This is done to make intensities commensurably comparable from source to source, which is necessary if intensity patterns across sources are to be identified.

As used herein, a "discordant" observation is one that is " ... statistically unreasonbable [or extreme] on the basis of some prescribed probability model." [17]

As used herein, "exceptional" means a quantity that is markedly different from the other quantities against which it is compared.

As used herein, "selective expression" is defined as a pattern among a collection of intensities in which there is an intensity which is markedly elevated, or markedly depressed, against a baseline level of intensity characteristic of the collection of intensities being compared. Hence, a "selectively expressed" intensity is an exceptional intensity. In particular, selective expression is a pattern in which there is a marked difference of intensity in a single source from a baseline level of expression established by the gene's or the entity's intensities in a source set. See Figure 1 for stereotypical examples. The method of the invention does not require,

however, that comparisons be made against all known sources. Instead, a carefully chosen subset of the known sources can be considered, especially since selective expression is a relative, not an absolute, assessment. Choice of source set enables the scientific context for expression comparisons to be tailored to the scientific questions being asked: organ systems vs. one another, tissues vs. one another (e.g., endothelium vs. smooth muscle or fibroblast), drug dose responses vs. one another, human vs. non-human species, chemical assays v. one another, etc.

A particular application of the invention provides a method that robustly identifies genes or proteins that are selectively expressed. The method combines assessments of the reliability of expression quantitation with a statistical test of intensity patterns. The method is applicable to small studies or to data mining of abundance data from large expression databases, whether mRNA or protein. The algorithm uniquely combines together a statistical test of discordancy, adjustments for baseline levels of the intensities (where baselines can be determined by source quality weighted averages), and adjustments for the separation of the largest and another intensity (gap) to give an overall assessment of confidence in selective expression. The algorithm achieves this by combining defined values -- baseline adjusted discordancy and gap -- into a decision function.

The algorithm is generally applicable to small- or large-scale expression-like data whether derived from DNA sequencing, proteomics, compound assays, pharmacogenomics, or toxicological safety assement, etc. The method can be implemented as computer programs that analyze databases of gene abundances on a regular basis.

The method is particularly useful in identifying biologically and pharmacologically interesting selectively expressed genes, hence, having objective implications for further analysis. It is well-established that DNA sequence copy number and mRNA levels in eukaryotic cells are present in a variety of abundance classes [1-3]. Very wide differences in gene expression level, i.e., in intracellular mRNA copy number, abundance, or in amount of gene product, are possible within the same cell. For example, it has been estimated that the copy numbers of expressed genes can vary from 1 to about 200,000 [4].

Further, the same cell type, as well as different cell types, may exhibit different patterns of gene expression when exposed to different conditions [5, 8]. Assessing differences in expression patterns, therefore, can be used to gauge differences in cell physiology and tissue behavior, intrinsically or in response to many different kinds of stimuli. As these differences may be correlated with fundamental biological phenomena or disease processes, delineations of patterns of gene or protein expression among normal and diseased states or patients exposed to drugs are of increasing importance in medical diagnostics and therapy.

Two stereotypical simple selective expression situations are possible: "up," where expression is significantly elevated in a specific tissue when compared against the baseline level in the other tissues; "down," where the expression in a specific tissue is significantly depressed when compared against the baseline expression in the other tissues. "Up" selective expression may be an important indication that the gene has been specifically activated, up-regulated, or its product differentially elevated in association with certain phenomena or agents affecting a particular tissue's biology. Similarly, "down" selective expression is either a significant down-regulation or essentially an inactivation of the gene (e.g., tumor suppressor loss of function) in association with specific biological events. Such broad phenomena as morphogenesis, differentiation, metabolic alteration, mutagenesis, bacterial and viral infection, physiological stress, disease, drugs and therapeutic interventions, etc., can manifest or cause selective expression effects.

For example, the method of the invention can compare relative levels of mRNA transcripts or relative levels of protein products. Despite the inherent difficulties in precisely measuring which mRNA species are translated and in what relative proportions, reliable enough information on expression levels can be obtained [5, 11, 14]. Moreover, the established experimental techniques of cDNA and EST sequencing, especially when employed on a large scale, can provide ESTs that can be combined computationally into assemblies [9]. Assemblies can be interpreted as putative expressed genes, though to widely varying levels of confidence in the assignments of assemblies to genes [12, 13]. Abundances of expressed genes or assemblies obtained from sampling are dependent upon the depth

of the sampling [15, 16] and contribute to inaccuracies in the computed intensities [13].

In one embodiment, the invention provides a computational method (algorithm) of identifying selectively expressed values in intensity data comprising analyzing statistical discordancy and gap criterion in a decision function wherein the decision function provides an overall confidence of above- or below-baseline exceptional intensity identification. The statistical discordancy can be adjusted for baseline intensity levels.

In an alternate embodiment, the invention provides a method of identifying exceptional values in intensity data comprising:

(a) selecting intensity values from intensity data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensities exceeds a predetermined minimum;

(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of exceptional intensity;

(f) identifying the degree of overall confidence of exceptional intensity; and

(g) displaying the results of step (f) on an output device.

In another embodiment, the invention provides a method of detecting selective expression of genes or gene products comprising:

(a) selecting intensity values from gene product data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensities exceeds a predetermined minimum;

(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of selective expression;

(f) identifying the degree of overall confidence of selective expression; and

(g) displaying the results of step (f) on an output device.

In these embodiments, the statistical discordancy test results of step (c) can be adjusted according to the difference between a baseline position and a maximum allowed intensity to achieve a baseline adjusted statistical significance. Preferably, the gap is determined between the largest and the next-to largest intensity. Further, when available, source quality confidence is based on trust, reliability, knowledge of error or relevance. Preferably, the intensity baseline position is determined by a source quality weighted average of the intensities.

In addition to display on an output device such as a monitor or a printer, the identity of the selectively expressed gene products can be stored in a database. The methods of the invention can further comprise the step of characterizing the selectively expressed gene product. Characterization can be done on the basis of of sequence, structure, biological function or other related characteristics. Once categorized, the database can be expanded with information linked to biological function, structure or other characteristics. Further, selectively expressed genes or gene products can be characterized on the basis of expert commentary from relevant human specialists or by the results of biological experiments. If desired, the selectively expressed entites detected by the method may be confirmed experimentally by techniques well known to those skilled in the art [2, 5-7].

In step (a), minimum source quality weight criterion are applied. For an entity's collection of intensities to be analyzed from the source set (e.g., a particular gene's abundances in a source set of libraries), intensities are selected from only those sources whose corresponding quality weight (i.e., trust, reliability, or relevance) exceeds a minimum. Minimum quality thresholds can be determined by those skilled in the art by applying scientific judgments concerning the reliabilities

or relevances of the sources. Oftentimes as data is being accumulated, a source's quality will change with the data, requiring the selective expression algorithm to be re-applied. Source quality weighting is considered optional, in which case this is equivalent to either no weighting or all weights being the same, e.g., unity.

Step (b) determines whether the number of selected intensity values exceeds a predetermined minimum. In sub-step (b1), there is the option of whether or not zero intensities in the source set are considered or ignored. If the option of ignoring, hence omitting, zero intensities is taken, then sub-step (b2) determines whether or not a non-zero intensity exceeds its source's detection limit (experimentally or computationally). In sub-step (b2) if a non-zero intensity does not exceed its source's detection limit, then that intensity is considered equivalent to zero and therefore omitted as in sub-step (b1). For an entity being analyzed for selective expression (e.g., a particular gene in a source set of libraries), if there is at least a predetermined minimum number of intensities surviving this step and that exceeds appropriate detection limits (discussed below), this entity (e.g., gene) and these intensities are marked for further analysis. In general, the minimum number of intensities will be enough to make confident identifications of exceptional intensities. However, a lesser number can be used with the understanding that the confidences in the assessments will be lower [17]. The minimum number of intensities is 3. Most preferably, the minimum number of intensities will be at least 10.

With respect to intensity detection limits, if an intensity appears to be absent from a particular source, then either (1) the intensity is actually not expressed in the source, or (2) the intensity is indeed expressed in the source but is smaller than the minimum intensity which can be measured, the detection limit. In case (2), since the intensity is not truly absent but instead occurs below the detection limit, it is thus recorded as absent. In the method of the invention, absent intensities can be considered as genuine absence only for very high quality sources with very low detection limits. All absent or sub-detection limit intensities are therefore ignored. However, the method does not require adopting this philosophy.

Step (c) applies a statistical discordancy test to identify statistically significant exceptional intensity values. Statistical tests of discordancy are known to

those skilled in the art [17-20]. The resulting statistical significance is used to score how exceptional the putative discordant intensity is. The test is applicable to exceptionally small intensities ("down" selective expression) as well as exceptionally large intensities ("up" selective expression).

A uniform distribution Dixon test [17] can be used in the method of the invention for the statistical test of discordancy. A uniform distribution assumes only that intensities are finite and there is no *a priori* most probable intensity. This is a reasonable parsimonious choice for an actually unknown inter-source intensity distribution; it is a choice which confers *a priori* only a very weak bias in distribution shape or in central tendency.

The first graph in Figure 1 diagrammatically shows a source set of intensities having a single exceptionally large intensity. Such data can be sorted in ascending order and re-plotted as in Figure 2. When values are sorted, the relative separation between the largest value and the remaining values becomes clearer. The size of the gap between the largest and next largest value divided by the distance between the largest and smallest values (see Figure 2) is an obvious measure of the separation of the largest value from all the other values. This "separation ratio" (equation 4 below) is the core of the statistic employed in the Dixon test for a single largest discordant value among uniform samples [17]. It captures the logical underpinnings of the statistical test.

In the case of the more general m^{th} largest discordant value Dixon test, the appropriate changes in the formulas for the degrees of freedom and the separation ratio dependent statistic [17] can be employed. The more general case is applicable to the problem of simultaneously identifying more than one selectively expressed intensity in a collection of intensities. For application of the test to selective expression, it was found that the single largest value test was sufficient and is preferred. The mathematical details follow.

For a selected entity (e.g., gene), let the vector $\mathbf{f}'$ comprise the entity's intensities from the n different sources of the source set which are to be analyzed after step (b). Let $\mathbf{q}$ be the vector comprising the corresponding source quality weights. If source quality weights are not assigned, the elements of $\mathbf{q}$ are set to

unity. The elements of **f'** and **q** are real numbers >0. The sequential order of the vectors' elements is arbitrary since the order of the sources in the source set can be arbitrary. However, once an order of sources is chosen, the elements of **f** and elements of **q** must appear in the same order since the respective correspondences between qualities and sources must be maintained.

Essentially the same method that is used for the identification of exceptionally large intensities, i.e., "up" selective expression, can be employed with minor modifications for the identification of exceptionally small intensities, i.e., "down" selective expression. Define vectors **f** and $\mathbf{f_{down}}$ from **f'** as follows:

$$\begin{cases} f_{max} = maximum(\mathbf{f}') \\ \mathbf{f} = \mathbf{f}' / f_{max}, & \text{"up" selective expression} \\ f_{down} = 1 - \mathbf{f}' / f_{max}, & \text{"down" selective expression} \end{cases} \quad (1)$$

Though the mathematical form of the method is unchanged by using $\mathbf{f_{down}}$ in place of **f**, identifying exceptionally small values is fundamentally, and practically, different from identifying exceptionally large values. This is because there can be intensities in **f** that are so minute (though still above a very small detection limit) as to be measurements indistinguishable from noise, making them useless as reliable values in a discordancy test. One way to remedy this difficulty is to restrict **f** to comprise only those values that are considerably larger than the detection limit. However, once equation 1 is used, the same baseline adjustment technique used for **f** (step (d)) can be applied to $\mathbf{f_{down}}$. Define **x** as the vector that comprises the n elements of **f** sorted in ascending order, i.e., $x_{i-1} \leq x_i$. Next, compute the Dixon critical statistic $T_{critical}$ from the elements of **x** (equations 3 through 5 below). Then use the Dixon test (equation 2 below) to compute the discordancy significance probability of the largest intensity among these intensities being compared. According to the Dixon test for a single largest outlier [17], the significance probability sp that the largest sample is discordant, i.e., exceptionally large, is given by

$$sp = Probability\,[t \geq T_{critical}] = 1 - \int_0^{T_{critical}} F_{2,2n-2}(z)\,dz \quad (2)$$

where t is a dummy variable which represents any possible value of $(n-2)\tau / (1-\tau)$ for fixed n, F is the standard statistical F-distribution with degrees of freedom 2 and (2n-2) [21], and where

$$gap = x_n - x_{n-1} \tag{3}$$

$$\tau = gap / (x_n - x_1) \text{ (the separation ratio)}, \tag{4}$$

$$T_{critical} = (n-2)\tau / (1-\tau). \tag{5}$$

The interpretation of significance probability, *sp*, is the natural one: the smaller the significance probability, the more exceptionally large is the largest value, x_n, when compared against all the other values of **x**. The significance probability given by the fundamental equation (2) can be reduced algebraically [17] to the very simple form

$$\log_{10}(sp) = (n-2)\log_{10}(1-\tau). \tag{6}$$

Equation 6 conveniently quantitates the theoretical statistical significance that the largest sample is exceptionally large. From equation 6, the significance probability decreases markedly as the separation ratio τ approaches 1. Moreover, this effect is stronger, the larger the sample size *n*. For a fixed sample separation ratio τ, the logarithm of the significance probability decreases linearly with the number of samples n since $\tau<1$ (equation 6).

Note that the conventional Dixon definition of the separation ratio τ effectively normalizes the separation between the largest and next-to-largest intensities by the range spanned by all the intensities being compared. This is what confers an apparent dynamic range indifference to the Dixon test. However, the effective dynamic range of the analyzed intensities with respect to a maximum allowed intensity is important to the method of the invention. The mathematical details of the adjustment made to the Dixon test to remedy the test's otherwise indifference to dynamic range is discussed in the step (d) details below.

Note that it can be shown numerically and analytically that

$\Delta \log(sp) \approx \frac{\partial \log(sp)}{\partial \tau} \Delta\tau \approx \frac{\partial \log(sp)}{\partial \tau} \Delta((x_n - x_{n-1})/(x_n - x_1))$ is small for small changes in *gap* or in any of x_1, x_{n-1}, or x_n. This obviates replacing any of x_1,

x_{n-1}, or x_n by respective source quality weighted estimates in the computation of τ in equation 4 above. However, a role for **q** persist in step (d). In step (d), the statistical discordancy test results are adjusted according to the difference between a baseline position and a maximum allowed intensity to achieve a baseline adjusted statistical significance. The baseline position can be determined by a source quality weighted average of the intensities. Apart from the putative discordant intensity, the other intensities among those being compared can be characterized as being clustered about a baseline level. The statistical test of discordancy results from step (c) are adjusted according to the difference between the baseline position and the maximum allowed intensity. The adjustment to the statistical significance is to increasingly downgrade it as the baseline becomes closer to the maximum allowed intensity. The baseline dependent adjustment is based on the dynamic range of the values being increasingly compressed, hence less mutually distinguishable, the closer the baseline is to the allowed upper limit. But, the Dixon test is indifferent to dynamic range compression, as noted above. However, since the discrimination of values is necessarily eroded as the effective dynamic range is compressed, the confidence in outlier detection (discordancy) should be eroded correspondingly. The mathematical details are explained below.

The position of the baseline, i.e., a level which characterizes the non-extreme values of a collection of intensities, should affect the confidence of the selective expression determination as described above. Along these lines, if the dynamic range is compressed in the extreme, then the measurements would all become essentially indistinguishable since the accuracy of real measurements is always limited. Hence, discordancy detection would be meaningless in such a situation, regardless of how discordancy is computed, since separations between the values involved would be indistinguishable from numerical or measurement noise. However, the Dixon test is indifferent to the dynamic range of the data, as noted in step (c). This phenomenon of indifference to dynamic range is not idiosyncratic to Dixon tests, but is inherent generally to any excess/spread, range/spread, or deviation/spread discordancy statistical test [17]. So, even if the dynamic range is compressed, as long as the difference between the largest and the next-to-largest

values is proportionally compressed, the traditional Dixon test significance is unchanged. Thus, the traditional Dixon test must be modified to correct for erosion in confidence in discordancy detection as a compression in dynamic range occurs.

To accomplish this, the Dixon significance is adjusted by a baseline adjustment factor λ. $\lambda \in (0,1)$ is designed to attenuate the traditional Dixon separation ratio τ (equation 4) so that the adjusted τ is

$$\tau_{adjusted} = \lambda\tau. \tag{7}$$

We choose λ to be a sigmoidal function of baseline with the parameters of the sigmoid chosen so that λ remains approximately unity until the baseline encroaches substantially on the maximum allowed intensity, e.g., typically 1. For example,

$$\lambda = \left(1 + \left(\frac{\hat{x}_{baseline}}{c}\right)^{b}\right)^{-1} \tag{8}$$

where c is the value of $\hat{x}_{baseline}$ for which $\lambda = 0.5$, i.e., the sigmoid's point of inflection, and $b > 0$ controls the steepness of λ decay with increasing $\hat{x}_{baseline}$. In practice, we typically use $c = 0.8$ and $b = 10$ in equation 8. $\hat{x}_{baseline}$ is a source quality weighted estimator of **x** baseline, which excludes the putative extreme value x_n, e.g., a weighted average

$$\hat{x}_{baseline} = \sum_{i=1}^{k} q_i x_i \Big/ \sum_{i=1}^{k} q_i \tag{9}$$

In equation 9, $k < n$ to insulate the baseline estimate from possible undue influence of a putative extreme value x_n. Though we prefer quality weighted baseline estimates, one can choose to ignore quality differences in $\hat{x}_{baseline}$, and therefore, substitute unity for the q_i. In which case, equation 9 becomes the simple average.

For this τ adjustment for baseline concept, any function can be chosen which has the effect of substantially diminishing outlier significance when baselines encroach upon the maximum allowed intensity. We find sigmoids to be especially convenient. Thus, the traditional Dixon outlier significance probability (equation 6) is adjusted for the baseline by the simple formula:

$$\log\left(sp_{adjusted}\right) = (n-2)\log(1-\tau_{adjusted}) \tag{10}$$

where $\tau_{adjusted} = \lambda\tau$, λ is computed from equations 7 and 8.

To illustrate, consider the examples in Fig. 3 and the corresponding Table 2. Each row in Table 2 represents a different, yet related, set of intensities. **x** denotes the vector comprising a set of intensities sorted in ascending order. In each example and throughout the calculations, the source set size is held constant at n = 22, and the maximum intensity x_n is held constant at 1. However, for each example (row) the minimum intensity x_1 is set to the value in the first column. For illustrative simplicity, x_1 is also taken to be the baseline estimate $\hat{x}_{baseline}$ since the non-extreme values are so narrowly clustered near x_1 in these examples. Quality weights are not needed, then, in these simplified baseline estimates.

TABLE 2 - Affect of Baseline Position on the Adjusted Dixon Statistical Significance Probability

Base-line	x_{n-1}	gap	λ	$\tau_{adjusted} = \lambda\tau$	$\log_{10}(sp_{adj.})$	$\Delta \log_{10}(sp)$
0.25	0.32	0.68	1.00	0.90	-20.00	0.00
0.50	0.55	0.45	0.99	0.89	-19.32	0.68
0.75	0.78	0.22	0.66	0.59	-7.75	12.25
0.90	0.91	0.09	0.24	0.21	-2.07	17.93

Each example set of synthetic intensity values corresponding to $\hat{x}_{baseline}$ values {0.25, 0.5, 0.75, 0.9} are plotted respectively in Fig. 3. In each case, the traditional Dixon significance probability ($\log_{10}$(sp) = -20) is kept fixed. Constant Dixon significance, regardless of baseline position, is achieved deliberately in these synthetic data by adjusting the second-largest intensity (x_{n-1}), shown in column 2, according to equations 3 and 7. Hence, the gap between the largest and next-to-

largest intensities ($x_n - x_{n-1}$) necessarily decreases as the baseline increases; yet, the traditional Dixon significance remains unchanged. But, the closer the baseline is to the allowed maximum, ($x_n = 1$), the less confidence there is in an assessment of discordancy. Therefore, the statistical significance must be reduced from the traditional Dixon value according to how the baseline encroaches upon the allowed maximum. This is done by diminishing the separation ratio τ according to a sigmoidal function of the baseline (equations 7 and 8). As can be seen, the baseline adjusted significance decreases as the baseline increases towards the allowed maximum The erosion of traditional Dixon significance increases as baselines are continuously increased towards the allowed maximum (Fig. 4). See also Table 2 where x_{n-1} (column 2) is computed by using equations 4, 5 and 6 to insure that the traditional Dixon discordancy significance probability remains fixed at $\log_{10}(sp) = -20$ even though x_1 is different in each example. The baseline adjustment factor λ computed using equation 8 with b = 10 and c = 0.8 is in column 4. The effect of the baseline adjustment factor λ on the traditional Dixon significance is shown in columns 5 and 7. The loss of statistical significance, $\Delta\log_{10}(sp)$, between the baseline adjusted significance and the traditional significance in column 7 is in $\log_{10}$ units. It is plotted as a continuous function of baseline in Fig. 4. As desired for baseline adjustments of statistical significance, the erosion in confidence reflected becomes substantial as the baseline encroaches upon an intensity upper limit.

An important general principle is illustrated by these examples: Though the traditional Dixon significance probability can remain apparently extremely significant (e.g., 10^{-20}) even as the dynamic range of the data is compressed ever smaller (represented here by the baseline coming ever closer to an allowed maximum), a baseline adjusted significance probability can nonetheless reflect the erosions of statistical significance that should occur in data whose dynamic range is substantially compressed.

It should be noted that while there is no intrinsic method to determine how much discordancy significance probability ought to be attenuated quantitatively as a function of baseline levels, scientific judgment of those skilled in the art concerning data accuracy, the resolving power of intensity measurement techniques, and the

dynamic range of intensity data can be used to design significance adjustment functions. The role of scientific judgment in this situation is analogous to that for establishing source quality weighting and for subjectively interpreting discordancy.

In step (d), a gap is determined by applying a minimum intensity gap criterion to the results of the statistical discordancy test. The gap, i.e., the separation between the largest and the next-to-largest intensities, is a fundamental ingredient in discordancy assessment. See Figure 2 and the description of step (c) above. If the gap is below or near the resolving power of the technique providing the intensity data, there is necessarily negligible confidence in the assessment of discordancy, regardless of how the discordancy statistical significance is computed. This is because a gap commensurable with the intensity measurement technique's resolving power means that the difference between the values constituting the gap is indistinguishable from measurement noise. Therefore, a minimum gap criterion should be applied in conjunction with the discordancy statistical test from step (c). While there is no objective formula for establishing the minimum gap criterion, scientific judgment of those skilled in the art can be used to set the minimum gap threshold which takes into account the accuracy and resolving power of the technique that provides the intensity data. The mathematical details of step (d) follow.

Those gaps which meet a minimum gap threshold g_{thresh} are rescaled linearly between g_{thresh} and the maximum allowed intensity x_{sup}. Call these rescaled gaps g, e.g:

$$g = \begin{cases} 0, \text{ if } gap \le g_{thresh} \\ (gap - g_{thresh})/(1 - g_{thresh}), \text{ if } gap > g_{thresh} \end{cases} \quad (11)$$

Analogously, linearly transform the baseline adjusted significance $\log_{10}(sp_{adjusted})$ (equation 10) between the weakest-to-strongest statistical significance that one is willing to accept, i.e., between $\log_{10}(sp)_{thresh}$ and $\log_{10}(sp)_{inf}$, respectively.The lower bound $\log_{10}(sp)_{inf}$ is the statistical significance

beyond which stronger statistical significance is essentially inconsequential. Denoting by s, ($0 \le s \le 1$), as this transformation gives:

$$s = \begin{cases} 0, \text{ if } \log_{10}(sp_{adjusted}) \ge \log_{10}(sp)_{thresh} \\ 1, \text{ if } \log_{10}(sp_{adjusted}) \le \log_{10}(sp)_{inf} \\ \dfrac{\log_{10}(sp_{adjusted}) - \log_{10}(sp)_{thresh}}{\log_{10}(sp)_{inf} - \log_{10}(sp)_{thresh}}, \text{ if } \log_{10}(sp)_{thresh} < \log_{10}(sp_{adjusted}) < \log_{10}(sp)_{inf} \end{cases}$$

(12)

Preferably, $\log_{10}(sp)_{thresh} = -5$. Less preferred is $\log_{10}(sp)_{thresh} = -3$. Preferably, $\log_{10}(sp)_{inf} = -20$, which allows the adjusted significance probability a dynamic range of 10^{15}.

In step (e), a decision function is applied to the baseline adjusted statistical significance and the gap to determine an overall confidence of selective expression. In step (f), the degree of overall confidence of selective expression is identified.

The gap from step (d) should be combined with the baseline adjusted statistical significance of discordancy from step (c) in order to provide an overall confidence of selective expression. This is accomplished by applying a decision function that is dependent upon both of these. The decision function d ranks the assessment into Low (weak), Medium (moderate), or High (strong) confidence of selective expression. But, if either a minimum baseline adjusted discordancy significance was not met or a minimum gap was not exceeded, that entity and its set of intensities is marked as not exhibiting selective expression. The construction and employment of a representative decision function is described below.

While there is no intrinsic method to determine the mathematical forms of decision functions, there is practical utility in assigning overall confidences to separate weak from strong predictions of selective expression. An interpretation of the strength of a result is often for setting priorities for further analyses of the data and new experiments.

Decision function *d* near 0 is interpreted as very weak overall confidence, while *d* near 1 is very strong overall confidence in selective expression. *d* is designed to capture the following notions of confidence:

		scaled gap g	scaled gap g
		weak (g~0)	strong (g~1)
scaled sig. prob. s	weak (s~0)	weak (d~0)	strong (d~1)
scaled sig. prob. s	strong (s~1)	moderate (d~1)	strong (d~1)

d is (1) strong when both the baseline adjusted $\log_{10}$sp and the gap are strong (i.e., both s and g are near 1); (2) weak when both the $\log_{10}$sp and the gap are weak (i.e. s and g near 0); (3) moderate when the $\log_{10}$sp is strong but the gap is weak; (4) but strong nonetheless when the gap is strong yet the $\log_{10}$sp is weak. Notions (3) and (4) make sense because both the $\log_{10}$sp and the gap that are considered in the decision function confidence assessment are stronger their respective minimum thresholds. Either $\log_{10}$sp or gap weaker than their respective minimum thresholds is not selective expression, and immediately $d = 0$ in such cases. There is no *a priori* requirement that d be symmetrical with respect to g and s. In fact, in practice, an asymmetry is preferred that gives d near 1 for large gaps as long as $\log_{10}$sp is stronger than a threshold value. Using these principles, a useful decision function is:

$$d(g,s) = 1 - \left[(1-s)^{\alpha} (1-g)^{\beta} \left(\frac{\delta(1-g) + (1-\delta)(1-s)}{(1-g) + (1-s)} \right)^{\gamma} \right]^{\phi} \quad (13)$$

where $\alpha \geq 0$, $\beta \geq 0$, $\gamma \geq 0$, and δ ($0 < \delta < 1$) are independent parameters chosen empirically, and where ϕ is defined by $\phi = (\alpha + \beta + \gamma)^{-1}$. Observe that the term in brackets amounts to a numerical version of a logical AND of three terms, the third term of which amounting to a numerical logical OR of two terms blended in a proportion controlled by δ. Typically, we choose $\alpha = \beta = \gamma = 1.5$ and $\delta = 0.3$. Fig. 5 shows this decision function d plotted as a series of constant-d contours on (g,s)-

space. (g,s) are the respective linear transformations of gap and baseline adjusted $\log_{10}(sp)$ between the weak thresholds and strong limits. See equations 11-13.

Step (f): Though there is no intrinsic method for setting break points between weak, moderate, and strong overall confidences, in practice the strength of the selective expression overall degree of confidence breakpoints for *d* are taken to be 1/3 and 2/3, respectively.

Another aspect of the invention is a computer system for identifying selectively expressed values in intensity data. A representative computer system includes a hardware environment on which the methods of the invention may be implemented. The hardware environment includes a central processing unit, a memory device, a display and a user interface device. An exemplary hardware environment is a Sun Microsystems Ultra 1 running a UNIX operating system, having a display and keyboard and/or mouse input devices.

In one embodiment, the computer system for identifying selectively expressed values in intensity data comprises means for analyzing statistical discordancy and gap criterion in a decision function wherein the decision function provides an overall confidence of above- or below-baseline exceptional intensity identification.

In another embodiment, the computer system for identifying exceptional values in intensity data comprises:

(a) means for selecting intensity values from intensity data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) means for determining if the number of selected intensities exceeds a predetermined minimum;

(c) means for applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) means for determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) means for applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of exceptional intensity;

(f) means for identifying the degree of overall confidence of exceptional intensity; and

(g) means for displaying the results of step (f) on an output device.

In another embodiment, the computer system comprises a central processing unit executing a selectively expressed value identifying program stored in a memory device accessed by the central processing unit; a display on which the central processing unit displays screens of the exceptional value identifying program in response to user inputs; and a user interface device.

Another aspect of the invention is a computer readable medium containing program instructions for identifying selectively expressed values in intensity data comprising analyzing statistical discordancy and gap criterion in a decision function wherein the decision function provides an overall confidence of above- or below-baseline exceptional intensity identification.

In another embodiment, the computer readable medium contains program instructions for identifying exceptional values in intensity data, the program instructions comprising:

(a) selecting intensity values from intensity data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensities exceeds a predetermined minimum;

(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of exceptional intensity;

(f) identifying the degree of overall confidence of exceptional intensity; and

(g) displaying the results of step (f) on an output device.

The present invention will now be described with reference to the following specific, non-limiting examples.

Example 1

Selective Expression Detection in Synthetic Data

In Fig. 6, synthetic data representative of real assembly abundances are shown. Panel A shows Set 2 (filled circles) and Set 1 (open circles) for comparison; panel B shows Set 3 (filled circles) and Set 1 (open circles) for comparison. In panels A and B, the putative selective expression occurs in the third Source. Panel C shows the source qualities corresponding to the intensities.

The numerical values of the source qualities and corresponding intensity data are in Table 3. The computed numerical results using the method of the invention are summarized in Table 4. Though these intensity and source quality data are synthetic, they are representative of real data derived from a large database of gene abundances and library qualities.

TABLE 3 - Synthetic Intensity (Abundance) and Source (Library Quality) Assembly Data

Source	Quality	Example 1	Example 2	Example 3
1	0.26	0.19	0.35	0.64
2	0.27	0.29	0.39	0.68
3	0.22	0.92	0.71	1.00
4	0.20	0.24	0.37	0.66
5	0.26	0.37	0.43	0.72
6	0.65	0.31	0.40	0.69
7	0.29	0.21	0.35	0.64
8	0.26	0.10	0.30	0.59
9	0.26	0.30	0.40	0.69
10	0.26	0.23	0.37	0.65
11	0.21	0.35	0.43	0.72

12	0.28	0.22	0.36	0.65
13	0.26	0.21	0.36	0.64
14	0.25	0.26	0.38	0.67
15	0.22	0.17	0.34	0.63
$\hat{x}_{baseline}$		0.25	0.37	0.66
gap		0.55	0.28	0.28
• no baseline adjustment		0.67	0.68	0.68
• $_{adjust}$ baseline adjusted		0.67	0.68	0.58

TABLE 4 - Application of Selective Expression Algorithm to Synthetic Data

Set	Base-line Adjust	λ	*gap*	τ	$\log_{10}(sp)$	*d*	Comments
1a	no		0.55	0.67	-6.26	0.33	Reference example.
1b	yes	1.00	0.55	0.67	-6.27	0.33	Same as 1a; λ has no effect.
2a	no		0.28	0.68	-6.26	0.24	*d* different from 1a due to *gap* only.
2b	yes	0.99	0.28	0.68	-6.28	0.24	*d* different from 1a due to *gap*; λ has no effect.
3a	no		0.28	0.68	-6.26	0.24	*d* different from 1a due to *gap* only.
3b	yes	0.87	0.28	0.58	-4.90	0.00	*d* different from 1a due to $\lambda_{-adjusted}$ log10(*sp*)<-5, hence *d*=0.

To convey the effects of various components of the method, each Set 1, 2 and 3 of Fig. 6 and Table 3 was deliberately constructed to have very similar qualitative patterns of intensity vs. source. Yet, the examples are different in overall confidence of selective expression as determined by the method. In particular, each Set has the same source set (size n = 15) and, moreover, exactly the same separation ratio ($\tau = 0.67$) before any adjustments are made for baselines. Hence, these sets have by design exactly the same traditional Dixon significance probability before baseline adjustment. Table 4 columns display, respectively: the Set identification number corresponding to Fig. 6; whether a baseline adjustment was used in the discordancy computation (equation 7); baseline adjustment factor λ (equation 8), gap (equation 3), τ (equation 4 if no baseline adjustment, otherwise equation 7), discordancy significance probability $\log_{10}$sp (equation 6 or 10), decision function d (equation 13), and comments. Equation 9, which employs source qualities from Table 3, is used for the baseline estimates $\hat{x}_{baseline}$ in equation 8. The equation 8 sigmoidal parameters are b = 10 and c = 0.8. The parameter values in the decision function (equations 11-13) are $\alpha = \beta = \gamma = 1.5$, $\delta = 0.3$, $g_{thresh} = 0.25$, $\log_{10}(sp)_{thresh} = -5$, and $\log_{10}(sp)_{inf} = -20$.The effects of adjusting significance probability for baseline can be seen in Table 4 by comparing each Set's case b against its respective case a, which is unadjusted for baseline. Example 3b is the only one in which significance probability is non-negligibly changed by baseline adjustment. This can be appreciated by observing the effects of baseline on λ, hence on τ, when compared against the case 1a τ. Sets 2 and 3, however, have markedly smaller gaps than does Set 1. These diminutive gaps are responsible for the decision function values for Sets 2 and 3 being much smaller than for Set 1 even though the discordancy statistical significance probabilities (with or without baseline adjustments) are not changed much. The exception is case 3a, which has an ample loss of significance probability due to baseline adjustment. Though the 3b gap is the same as 3a, 3b's decision function is zero because baseline adjustment of its statistical significance probability has resulted in its $\log_{10}(sp)$ not meeting the minimum significance criterion $\log_{10}(sp)_{thresh} = -5$. Taken together, these

examples illustrate how qualitatively similar intensity vs. source patterns can have different overall confidence of selective expression (indicated by the decision function values), depending on the baseline of the data and the size of the gap, even when the expression patterns have essentially identical unadjusted traditional discordancy significance probabilities. By analyzing these examples, it can be seen how the qualitatively stronger confidence of selective expression of Set 1 as compared to Sets 2 and 3 (which is informally conveyed in Fig. 6) is quantitated through the decision function of the selective expression method applied to the data.

Example 2

Selective Expression Detection in Gene Expression Data

To convey the appearances of stereotypical selective expression patterns in real gene expression data, intensity vs. source plots of some actual examples of algorithmically identified Extremely Strong, Strong, and Weak overall confidence selective gene expression are shown in Fig. 7, panels A, B, and C, respectively. Shown are intensity (abundance) vs. source (library) plots for three actual assemblies from a database of real sources and assembly abundances. Assembly A has a extremely strong overall confidence of selective expression (decision function d = 1.0). Assembly B has a strong overall confidence of selective expression (d = 0.75). Assembly C has weak overall confidence of selective expression (d = 0.31). Summarized algorithmic calculations corresponding to these examples are displayed in Table 5. The columns are similar to those in Table 4. In these particular real examples, baseline adjustment has no effect since the baselines are well below 0.8. Hence, the discordancy statistical significance probabilities are the same as the unadjusted statistical significances.

It is easily determined visually from the plots in Fig. 7 that the τ are decreasing from example A to C, with the larger decrease being from B to C. The corresponding τ are actually $\{0.78, 0.67, 0.35\}$, which agrees with this qualitative visual observation. That the discordancy statistical significance probabilities increase so dramatically with this series of τ values is due to the considerable size of the n involved, $\{87, 41, 49\}$, respectively. The marked difference in $\log_{10}(sp)$ between A and B is much more due to the difference in n

than in τ. However, the substantial difference in $\log_{10}(sp)$ between B and C is due to the difference in τ more than the difference in n. These quantitations are not surprising given equation 6. Clearly, A exhibits maximum confidence as can be seen visually in Fig. 7 and quantitatively in Table 5. That the d for C is half that for B is due to both the gap and the $\log_{10}$(sp) in combination being weaker in C than B.

TABLE 5 - Selective Expression in Gene Expression Data

Set	n	$\hat{x}_{baseline}$	λ	gap	τ	$\log_{10}(sp_{adj.})$	d	Overall Confidence in S. E.
A	87	0.03	1.0	0.78	0.78	-56.0	1.0	Very Strong
B	41	0.10	1.0	0.66	0.67	-18.8	0.7	Strong
C	47	0.20	1.0	0.34	0.34	-8.5	0.3	Weak

While it is useful for better understanding the data to dissect the various relative contributions of the ingredients of the selective expression algorithm as done above, the real power of the decision function *d*, is its utility in qualitatively ranking overall confidence in selective expression patterns in large scale data in a way that is not only easily automated, but objective and consistent.

References

All publications from the scientific literature cited in this specification are herein incorporated by reference as though fully set forth.

[1] R. J. Britten and D. E. Kohen, "Repeated Sequences in DNA.," Science, vol. 161, pp. 529-540, 1968.

[2] G. A. Galau, W. H. Klein, R. J. Britten, and E. H. Davidson, "Significance of Rare mRNA Sequences in Liver," Archives of Biochemistry and Biophysics, vol. 179, pp. 584-599, 1977.

[3] B. D. Hames and S. J. Higgins, "Nucleic Acid Hybridisation — A Practical Approach," in The Practical Approach Series. Oxford, UK: IRL Press Limited, 1985, pp. 245.

[4] S. Patanjali, S. Parimoo, and S. M. Weissman, "Construction of a Uniform-Abundance (Normalized) cDNA Library," Proceedings of the National Academy of Sciences USA, vol. 88, pp. 1943-1947, 1991.

[5] M. D. Adams, "Expressed Sequence Tags as Tools for Physiology and Genomics," in Automated DNA Sequencing and Analysis, M. D. Adams, C. Fields, and J. C. Venter, Eds. London: Academic Press Ltd., 1994, pp. 71-80.

[6] M. Singer and P. Berg, Genes & Genomes. Mill Valley, CA: University Science Books, 1991.

[7] M. R. Wilkins, K. L. Williams, R. D. Appel, and D. F. Hochstrasser, "Proteome Research: New Frontiers in Functional Genomics," in Principles and Practice. Berlin: Springer-Verlag, 1997, pp. 243.

[8] H. Lodish, D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell, Molecular Cell Biology, Third Edition ed. New York: Scientific American Books / W. H. Freeman and Co., 1995.

[9] M. D. Adams, C. Fields, and J. C. Venter, "Automated DNA Sequencing and Analysis," London: Academic Press Ltd., 1994, pp. 368.

[10] N. L. Anderson, J.-P. Hofmann, A. Gemmell, and J. Taylor, "Global Approaches to Quantitative Analysis of Gene-Expression Patterns Observed by

Two-Dimensional Gel Electrophoresis," Clinical Chemistry, vol. 30, pp. 2031-2036, 1984.

[11] L. Anderson and J. Seilhamer, "A Comparison of Selected mRNA and Protein Abundances in Human Liver," Electrophoresis, vol. 18, pp. 533-537, 1997.

[12] C. Burks, M. L. Engle, S. Forrest, R. J. Parsons, C. A. Soderlund, and P. E. Stolorz, "Stochastic Optimization Tools for Genomic Sequence Assembly," in Automated DNA Sequencing and Analysis, M. D. Adams, C. Fields, and J. C. Venter, Eds. London: Academic Press Ltd., 1994, pp. 250-259.

[13] E. W. Myers, "Advances in Sequence Assembly," in Automated DNA Sequencing and Analysis, M. D. Adams, C. Fields, and J. C. Venter, Eds. London: Academic Press Ltd., 1994, pp. 231-248.

[14] B. R. Herbert, J.-C. Sanchex, and L. Bini, "Two-Dimensional Electrophoresis: The State of the Art and Future Directions," in Proteome Research: New Frontiers in Functional Genomics, M. R. Wilkins, K. L. Williams, R. D. Appel, and D. F. Hochstrasser, Eds. Berlin: Springer-Verlag, 1997, pp. 13-33.

[15] J. Bunge and M. Fitzpatrick, "Estimating the Number of Species: A Review," Journal of American Statistical Association, vol. 88, pp. 364-373, 1993.

[16] W. A. Lewins and D. N. Joanes, "Bayesian Estimation of the Number of Species," Biometrics, vol. 40, pp. 323-328, 1984.

[17] V. Barnett and T. Lewis, Outliers in Statistical Data: Chichester & New York, 1978.

[18] G. L. Tietjen, "The Analysis and Detection of Outliers," in Goodness-of-Fit Techniques, vol. 68, Statistics, Textbooks and Monographs, R. B. D'Agostino and M. A. Stephens, Eds. New York: Marcel Dekker, Inc., 1986, pp. 497-521.

[19] D. M. Hawkins, Identification of Outliers. London & New York: Chapman & Hall, 1980.

[20] R. B. D'Agostino and M. A. Stephens, "Goodness-of-Fit Techniques," in Statistics, Textbooks and Monographs, vol. 68. New York: Marcel Dekker, Inc., 1986.

[21] L. Sachs, Applied Statistics - A Handbook of Techniques, 2nd ed. New York: Springer-Verlag, 1982.

It is contemplated that other statistical tests of outlier discordancy may be used in place of the Dixon test [17] in Steps (c), (d), and (f). Further, the decision function may have a mathematical form different than equation (13) which may be used in Steps (f) and (g). The properties of a decision function *d* are what matters more than the particular mathematical form (e.g, equation (13)) that is chosen: Decision function *d* near 0 is interpreted as very weak overall confidence, while *d* near 1 is very strong overall confidence in selective expression. *d* is designed to capture the following notions of confidence:

		scaled gap *g*	**scaled gap *g***
		weak (g~0)	strong (g~1)
scaled sig. prob. *s*	weak (s~0)	weak (d~0)	strong (d~1)
scaled sig. prob. *s*	strong (s~1)	moderate (d~1)	strong (d~1)

d is (1) strong when both the baseline adjusted $\log_{10}$sp and the gap are strong (i.e., both s and g are near 1); (2) weak when both the $\log_{10}$sp and the gap are weak (i.e. s and g near 0); (3) moderate when the $\log_{10}$sp is strong but the gap is weak; (4) but strong nonetheless when the gap is strong yet the $\log_{10}$sp is weak. Notions (3) and (4) make sense because both the $\log_{10}$sp and the gap that are considered in the decision function confidence assessment are stronger their respective minimum thresholds. Either $\log_{10}$sp or gap weaker than their respective minimum thresholds is not selective expression, and immediately $d = 0$ in such cases. There is no *a priori* requirement that d be symmetrical with respect to g and s. In fact, in practice, an asymmetry is preferred that gives *d* near 1 for large gaps as long as $\log_{10}$sp is stronger than a threshold value.

It will be apparent to those skilled in the art that various modifications can be made to the present method without departing from the scope or spirit of the

invention, and it is intended that the present invention cover modifications and variations of the method provided they come within the scope of the appended claims and their equivalents.

Claims

1. A method of identifying selectively expressed values in intensity data comprising analyzing statistical discordancy and gap criterion in a decision function wherein the decision function provides an overall confidence of above- or below-baseline exceptional intensity identification.

2. The method of claim 1 wherein the statistical discordancy is adjusted for baseline intensity levels.

3. A method of identifying exceptional values in intensity data comprising:

(a) selecting intensity values from intensity data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensities exceeds a predetermined minimum;

(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of exceptional intensity;

(f) identifying the degree of overall confidence of exceptional intensity; and

(g) displaying the results of step (f) on an output device.

4. The method of claim 3 wherein the statistical discordancy test results of step (c) are adjusted according to the difference between a baseline position and a maximum allowed intensity to achieve a baseline adjusted statistical significance.

5. The method of claim 3 wherein the gap is determined between the largest and the next-to largest intensity.

6. The method of claim 1 or 3 wherein the intensity data is from tissue or cDNA library sources.

7. The method of claim 1 or 3 wherein the intensity data is from human sources.

8. The method of claim 1 or 3 wherein the intensity data is from non-human sources.

9. The method of claim 8 wherein the intensity data is from animal, plant, viral, bacterial, or microbial sources.

10. The method of claim 1 or 3 wherein the intensity data is from genomic sequencing, EST sequencing, microarray DNA hybridization, macromolecular gridding, compound assays, molecular screening assays, patient diagnostic or toxicological data sources.

11. The method of claim 3 wherein the source quality confidence is based on trust, reliability, knowledge of error or relevance.

12. The method of claim 3 wherein the intensity baseline position is determined by a source quality weighted average of the intensities.

13. The method of claim 3 further comprising the step of characterizing the selectively expressed genes or gene products.

14. A method of detecting selective expression of genes or gene products comprising:

(a) selecting intensity values from gene product data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensities exceeds a predetermined minimum;

(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of selective expression;

(f) identifying the degree of overall confidence of selective expression; and

(g) displaying the results of step (f) on an output device.

15. The method of claim 14 wherein the statistical discordancy test results of step (c) are adjusted according to the difference between a baseline position and a maximum allowed intensity to achieve a baseline adjusted statistical significance.

16. The method of claim 14 wherein the source quality confidence is based on trust, reliability, knowledge of error or relevance.

17. The method of claim 14 wherein the baseline position is determined by a source quality weighted average of the intensities.

18. The method of claim 14 further comprising the step of characterizing the selectively expressed genes or gene products.

19. A computer system for identifying selectively expressed values in intensity data comprising means for analyzing statistical discordancy and gap criterion in a decision function wherein the decision function provides an overall confidence of above- or below-baseline exceptional intensity identification.

20. A computer system for identifying exceptional values in intensity data comprising:

(a) means for selecting intensity values from intensity data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) means for determining if the number of selected intensities exceeds a predetermined minimum;

(c) means for applying a statistical discordancy test to identify statistically significant exceptional intensity values;

(d) means for determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

(e) means for applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of exceptional intensity;

(f) means for identifying the degree of overall confidence of exceptional intensity; and

(g) means for displaying the results of step (f) on an output device.

21. A computer readable medium containing program instructions for identifying selectively expressed values in intensity data comprising analyzing statistical discordancy and gap criterion in a decision function wherein the decision function provides an overall confidence of above- or below-baseline exceptional intensity identification.

22. A computer readable medium containing program instructions for identifying exceptional values in intensity data, the program instructions comprising:

(a) selecting intensity values from intensity data sources, wherein confidence in source quality exceeds a predetermined minimum threshold;

(b) determining if the number of selected intensities exceeds a predetermined minimum;

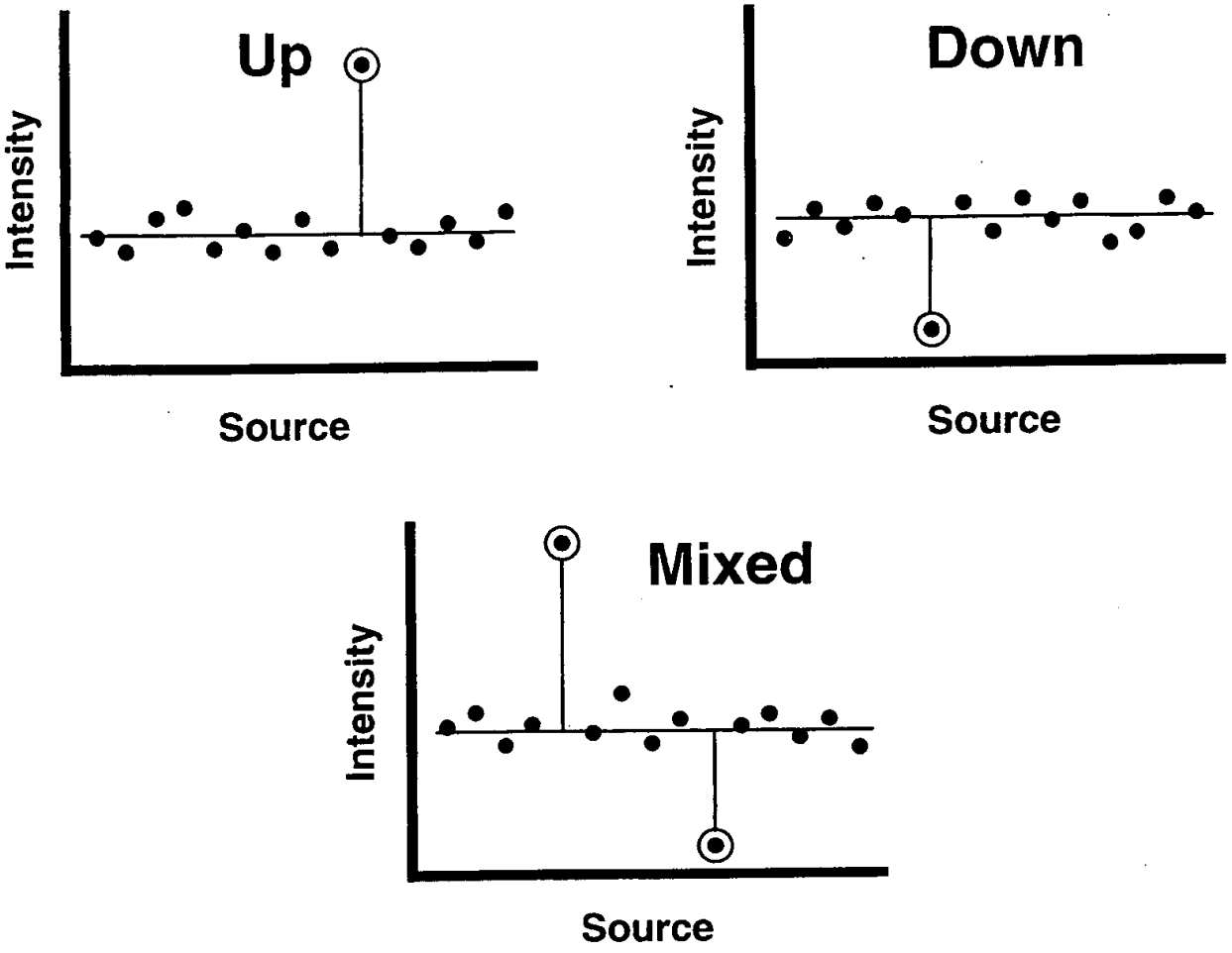
(c) applying a statistical discordancy test to identify statistically significant exceptional intensity values;

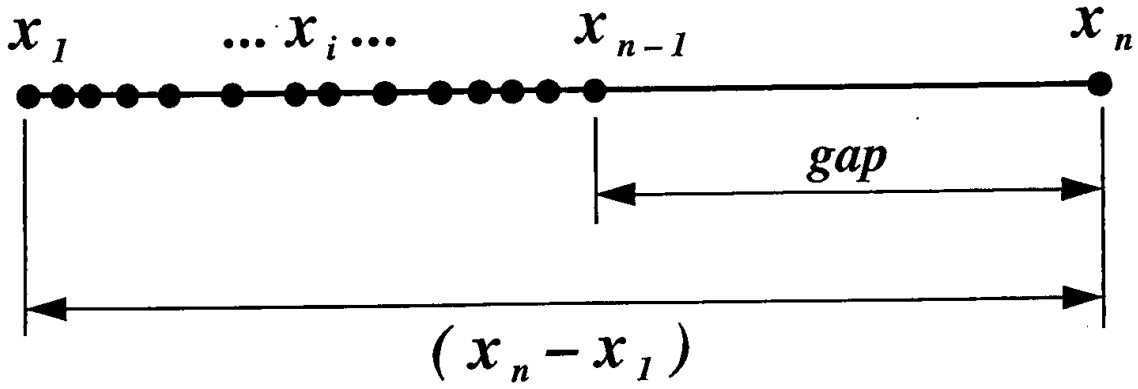
(d) determining a gap between the largest and another intensity by applying a minimum intensity gap criterion to the results of the statistical discordancy test;

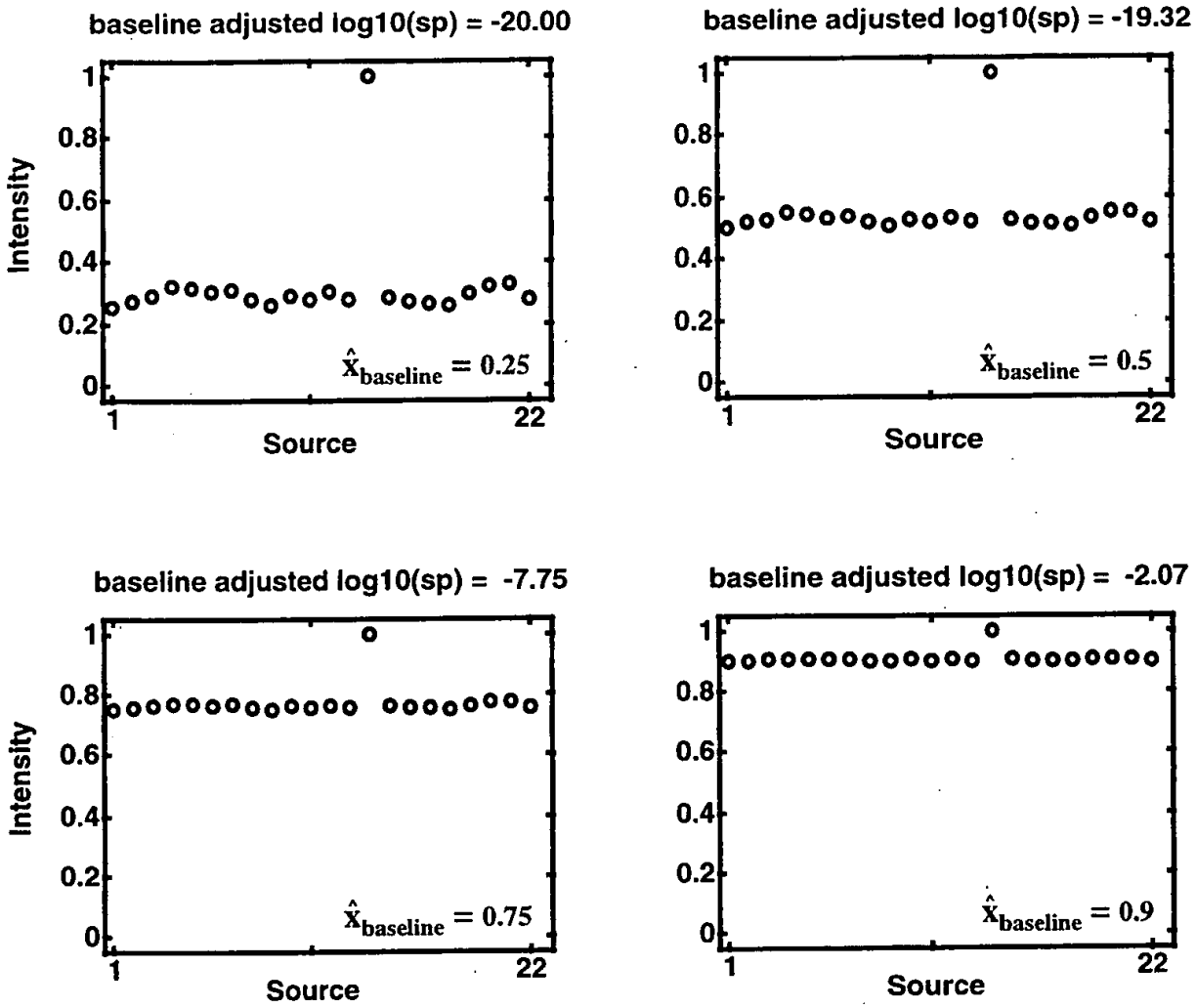
(e) applying a decision function to the discordancy statistical significance and the gap to determine an overall confidence of exceptional intensity;

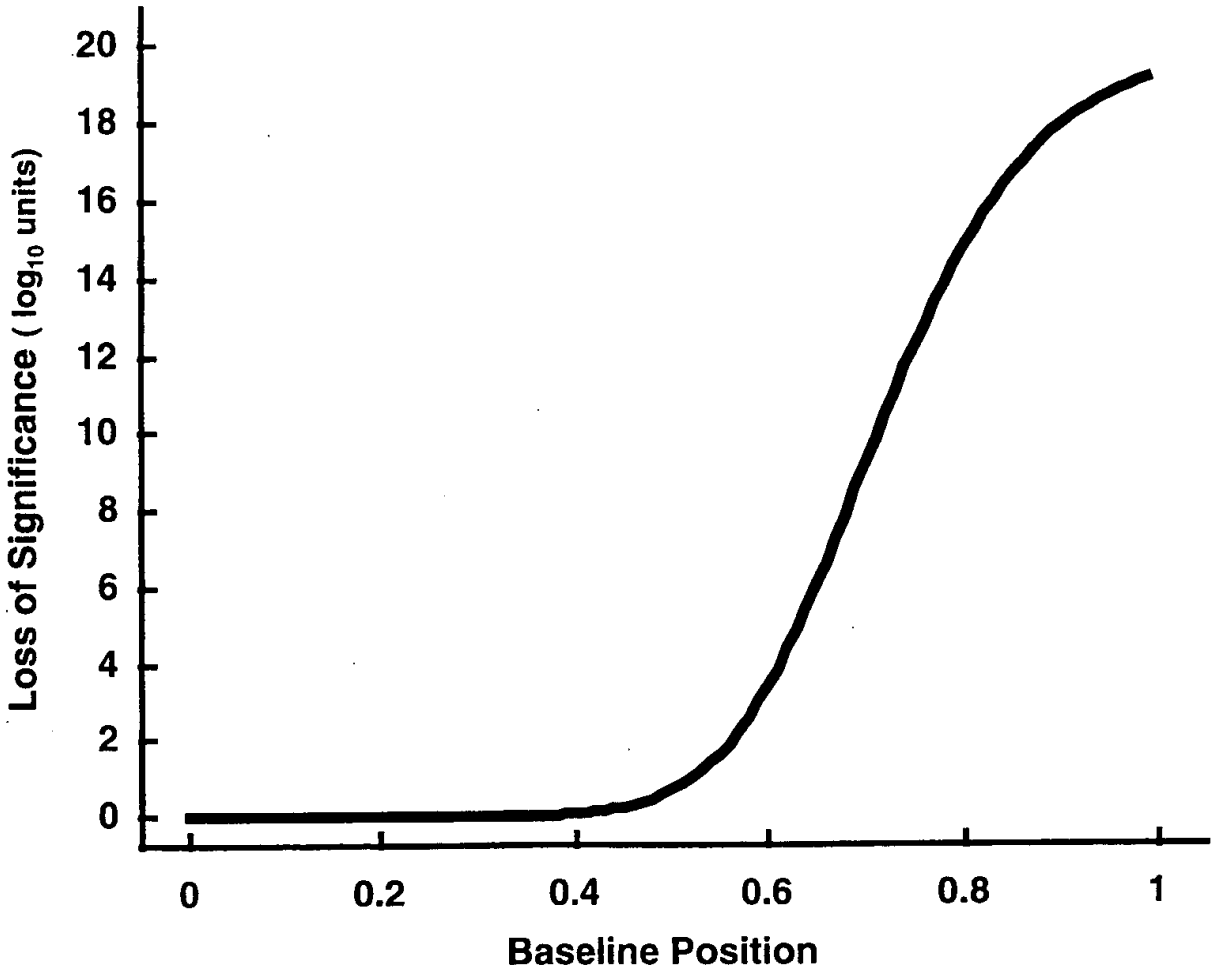
(f) identifying the degree of overall confidence of exceptional intensity; and

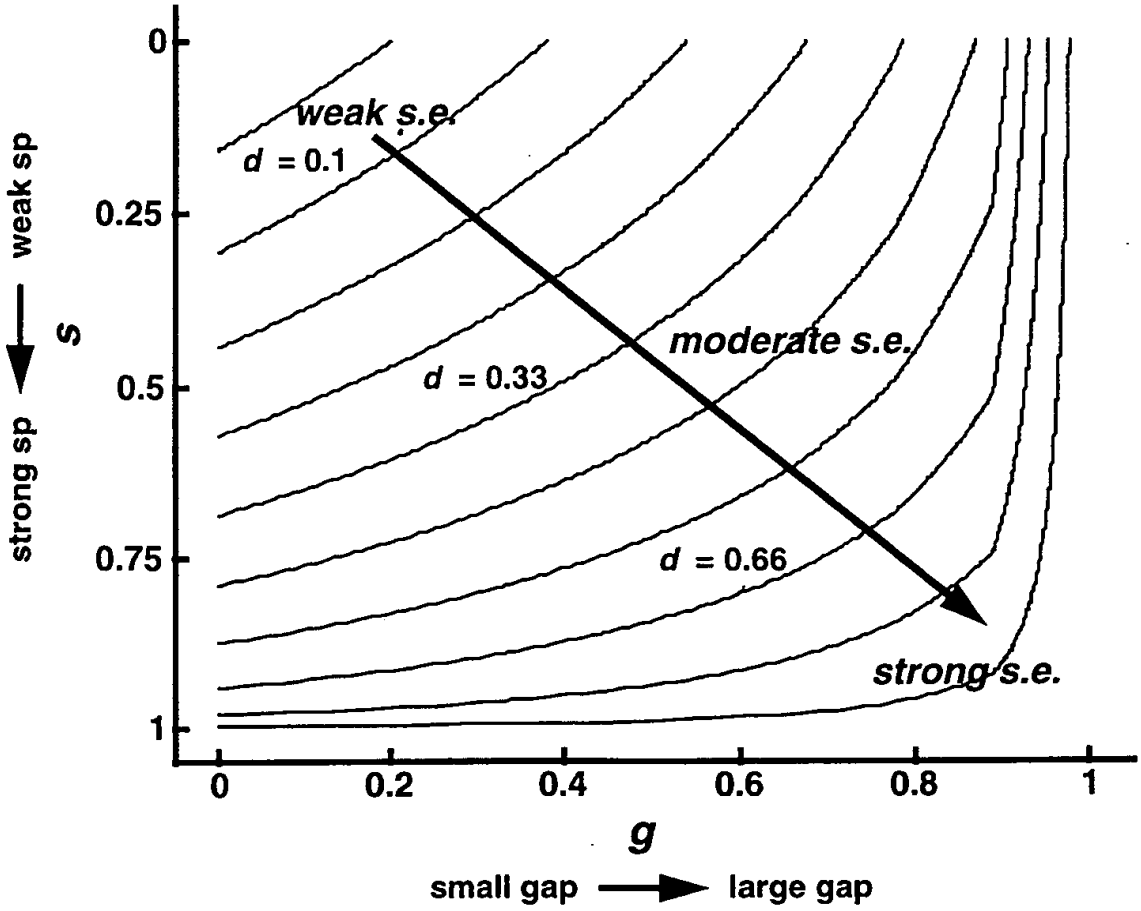
(g) displaying the results of step (f) on an output device.

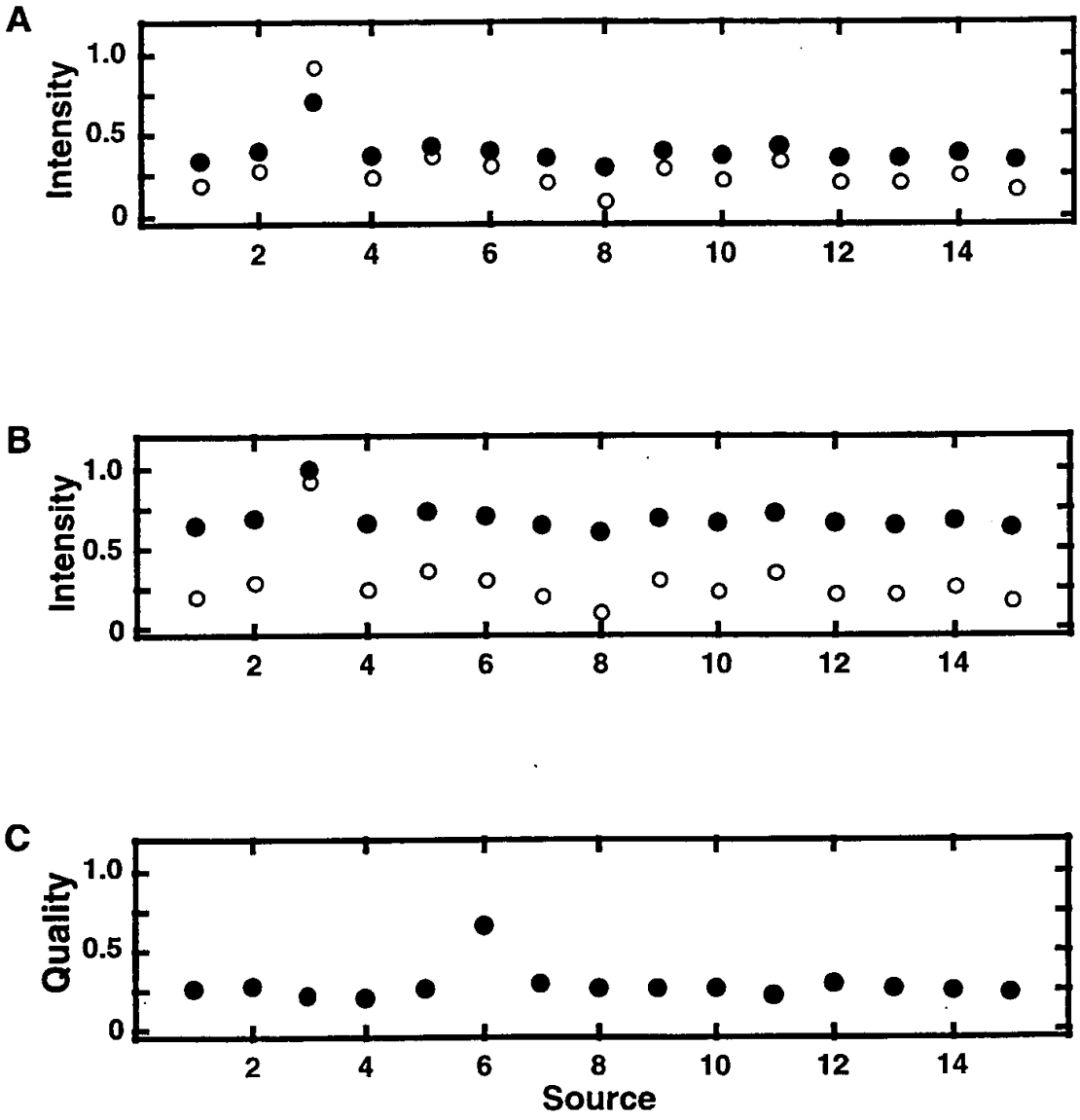

Up
Intensity
Source
Down
Intensity
Source
Mixed
Intensity
Source

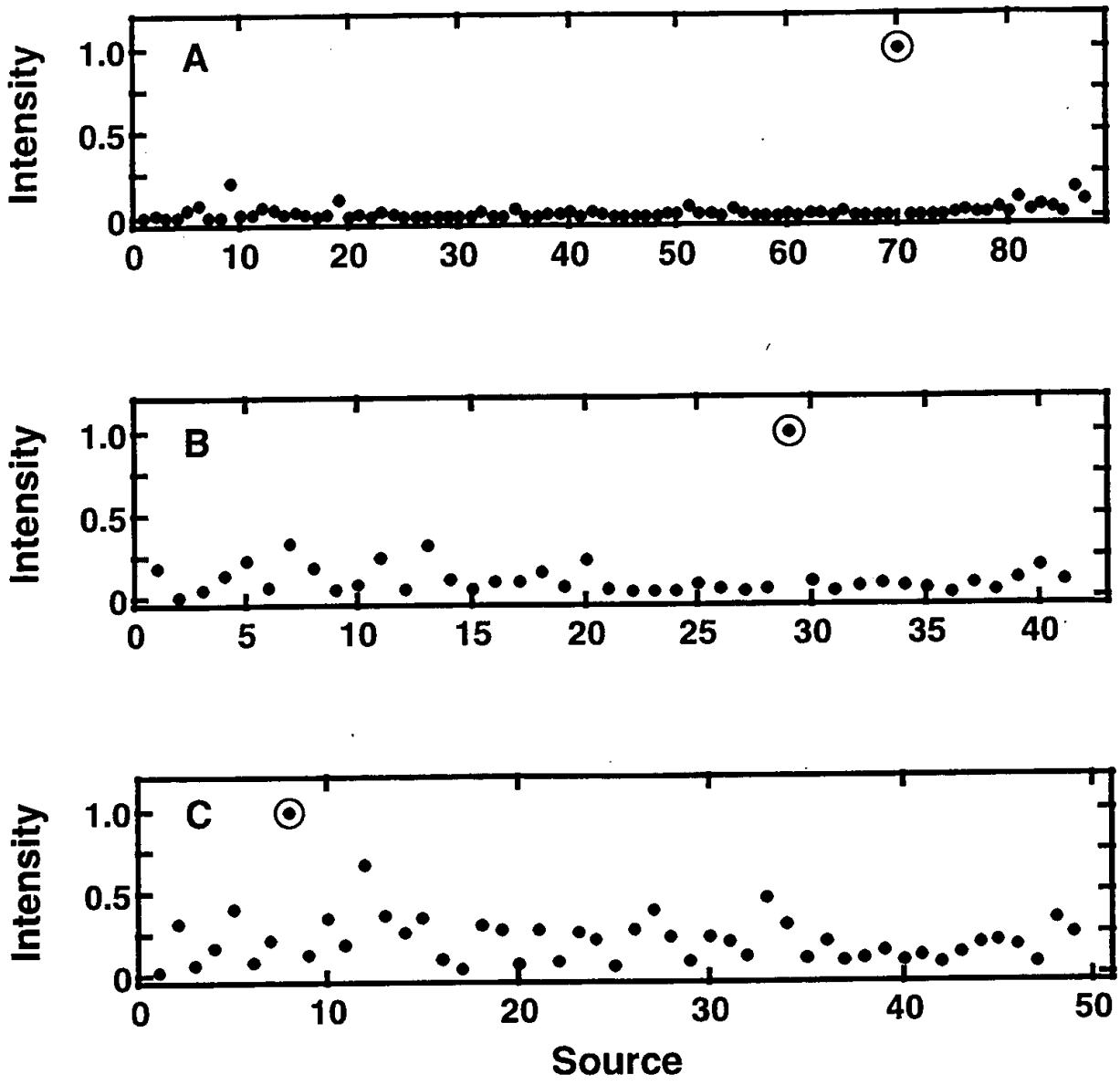

x_1
... x_i ...
x_{n-1}
x_n
gap
$(x_n - x_1)$


baseline adjusted log10(sp) = -20.00
baseline adjusted log10(sp) = -19.32
baseline adjusted log10(sp) = -7.75
baseline adjusted log10(sp) = -2.07
Intensity
Source
$\hat{x}_{baseline} = 0.25$
$\hat{x}_{baseline} = 0.5$
$\hat{x}_{baseline} = 0.75$
$\hat{x}_{baseline} = 0.9$
0
0.2
0.4
0.6
0.8
1
22


Loss of Significance (log10 units)
20
18
16
14
12
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
Baseline Position


0
0.25
0.5
0.75
1
s
strong sp
weak sp
weak s.e.
d = 0.1
moderate s.e.
d = 0.33
d = 0.66
strong s.e.
0
0.2
0.4
0.6
0.8
1
g
small gap
large gap


A
Intensity
1.0
0.5
0
2
4
6
8
10
12
14
B
Intensity
1.0
0.5
0
2
4
6
8
10
12
14
C
Quality
1.0
0.5
0
2
4
6
8
10
12
14
Source


A
B
C
Intensity
Source
1.0
0.5
0
0 10 20 30 40 50 60 70 80
0 5 10 15 20 25 30 35 40
0 10 20 30 40 50

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/11259

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G05B 5/34
US CL : 364/130

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 364/130

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, CAS, DIALOG

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	LEFEVRE et al. A fast word search alrithim for the representation of sequence similarity in genomic DNA. Nucleic Acid Research. 11 February 1994, Vol 22, Number 3, pages 404-411. see entire article.	1-7, 14-18
Y	LEFEVRE et al. Pattern recognition in DNA sequences and its application to consesus foot-printing. CABIOS. June 1993, Vol 9, Number 3, pages 349-354, see entire article.	1-7, 19-22
Y	US 5,214,717 (KIMURA et al) 25 May 1993, see entire article.	19-22
Y	US 5,068,909 (RUTHERFORD et al) 26 November 1991, see entire document	19-22

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

11 AUGUST 1999

Date of mailing of the international search report

21 OCT 1999

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

JEFFREY S. LUNDGREN

Telephone No. (703) 308-0196